Computational Workflow for Accelerated Molecular Design Using Quantum Chemical Simulations and Deep Learning Models

Andrew E. Blanchard[1], Pei Zhang[1],
Debsindhu Bhowmik[1], Kshitij Mehta[2], John Gounley[1], Samuel Temple Reeve[1],
Stephan Irle[1], and Massimiliano Lupo Pasini[1]

[1] Oak Ridge National Laboratory, Computational Sciences and Engineering Division,
Oak Ridge, TN 37831, USA, email: irles@ornl.gov
[2] Oak Ridge National Laboratory, Computer Science and Mathematics Division,
Oak Ridge, TN 37831, USA

Abstract. Efficient methods for searching the chemical space of molecular compounds are needed to automate and accelerate the design of new functional molecules such as pharmaceuticals. Given the high cost in both resources and time for experimental efforts, computational approaches play a key role in guiding the selection of promising molecules for further investigation. Here, we construct a workflow to accelerate design by combining approximate quantum chemical methods [i.e. density-functional tight-binding (DFTB)], a graph convolutional neural network (GCNN) surrogate model for chemical property prediction, and a masked language model (MLM) for molecule generation. Property data from the DFTB calculations are used to train the surrogate model; the surrogate model is used to score candidates generated by the MLM. The surrogate reduces computation time by orders of magnitude compared to the DFTB calculations, enabling an increased search of chemical space. Furthermore, the MLM generates a diverse set of chemical modifications based on pre-training from a large compound library. We utilize the workflow to search for near-infrared photoactive molecules by minimizing the predicted HOMO-LUMO gap as the target property. Our results show that the workflow can generate optimized molecules outside of the original training set, which suggests that iterations of the workflow could be useful for searching vast chemical spaces in a wide range of design problems.

1 Introduction

The "design and perfection of atom- and energy-efficient synthesis of revolutionary new forms of matter with tailored properties" is one of five scientific grand challenges articulated in the Basic Energy Sciences Advisory Report on Directing Matter and Energy [1]. In chemical sciences, one of the most coveted and impactful targets is the ability to design molecular compounds with desirable properties such as biological activity for molecular therapeutics [2, 3] or particular photo-optical properties geared towards photovoltaic applications [4], or the development of molecular dyes [5] or biomarkers [6]. While significant progress has already been achieved in the field of machine learning-assisted computational drug discovery [7, 8], the use of artificial intelligence (AI) protocols for the design of photoactive molecules is still in its infancy [9]. This situation can be attributed in part to the fact that the prediction of photo-optical properties for a given molecular structure requires computationally expensive quantum chemical calculations, namely the computation of molecules in their ground and excited states [10]. The computational generation of sufficiently large databases containing molecular structure and their optical properties is therefore far more costly than the calculation of bioactivity, which is traditionally performed using computationally much cheaper empirical scoring or classical force field calculations of protein-ligand interactions [11].

A reasonable shortcut to predicting photo-optical molecular properties is to approximate electronic excitation energies with energy differences between molecular orbital (MO) energy levels [12]. Of particular interest here is the energy difference between the highest occupied MO (HOMO) and the lowest unoccupied MO (LUMO). This so-called "HOMO-LUMO" gap often correlates very well with the lowest-energy, and hence most accessible, excited state that a molecule typically adopts upon energy intake due to the absorption of photons. Thus, HOMO-LUMO gaps have recently become the target of AI-based approached to photoactive molecules [10, 12, 13]. However, as mentioned above, this approach is only a reasonable shortcut, since for the inverse design of molecular structures with desirable photo-optical properties the entire absorption and/or emission spectrum over the energy range of visible light is required [10]. *Ab initio* multireference wavefunction electronic structure methods such as CAS-PT2 [14] and NEVPT2 [15] are able to cover these energy ranges and provide a measure for absorption/emission intensity via the prediction of oscillator strengths, and therefore accurately predict molecular dye candidate photophysical properties. However, this capability comes with a substantial price: The calculation of an UV/Vis absorption spectrum for molecules containing only tens of atoms are impractical on standard laptop or even Linux-operated workstations, due to the enormous required computational effort and resources in CPU power and memory. Computationally less demanding methods are density functional theory (DFT)-based excited states time-dependent (TD)-DFT methods [16], with the approximate TD-density-functional tight-binding (TD-DFTB) method being one of the computationally most economical methods [17].

In this work we present a vision for a DFTB- and TD-DFTB-based computational workflow for the inverse design of molecules with desirable optical properties. As others before us, we start with an approximation of excitation energies by MO energy differences [12, 13]. Besides the use of computationally efficient DFTB methods, our new inverse design algorithm has two major components: 1) Development of a multi-headed graph convolutional neural network (GCNN) approach that will allow the prediction of not only molecular orbital energy differences, but more complex properties such as multiple excitation energies and oscillator strengths for dye candidates [18, 19]. The GCNNs surrogate models are sufficiently fast to supplant computationally expensive, explicit excited state calculations required for the inverse design step, which will be component 2): Development of a machine learning masked language model (MLM)-based generation of new dye candidates with subsequent evaluation against a target function (desirable optical properties). As proof-of-principle, we present here a study for the inverse design of molecules with the lowest-possible HOMO-LUMO gap, which is motivated by possible applications as biomarkers in biomedical applications [20]. We note that our choice to minimize the HOMO-LUMO gap in this proposed workflow is arbitrary and could be replaced by a particular energy range. Our choice of a multi-headed GCNN surrogate will allow in subsequent works to extend our molecular design algorithm to predict novel molecules with optical spectra containing user-defined target features, such as intense optical absorption or emission in particular regions of the visible light spectrum.

2 Computational Workflow for Molecular Design

2.1 Inverse design of molecules with small HOMO-LUMO gap

It is important to remember that we are using two AI components in our computational workflow, which can be trained jointly or independently from each other. In this proof-of-principle work, we started with an existing generative MLM model that originally targeted the generation of drug candidates for molecular therapeutics [21], while the GCNN surrogate for the prediction of HOMO-LUMO gaps was specifically trained as part of this work.

Fig. 1 illustrates schematically a pipeline for the adaptive training of the GCNN surrogate model and the generation of new molecules with desirable HOMO-LUMO gap as predicted by the DFTB method. We begin by performing DFTB calculations on a subset of the 134,000 molecules contained in the QM9 database [22] to predict their HOMO-LUMO gaps, then train the GCNN surrogate which is subsequently utilized to score the newly generated molecules predicted by the MLM algorithm. In a final step, the newly predicted molecules are re-calculated using the DFTB method and their HOMO-LUMO gaps compared against the surrogate-predicted value (not shown in Fig. 1). In the following, we briefly discuss the center pieces of our workflow, namely the DFTB calculations of HOMO-LUMO gaps from SMILES strings, the GCNN surrogate and the MLM, and their interplay with each other, before elaborating the details of each component in subsequent Sections.

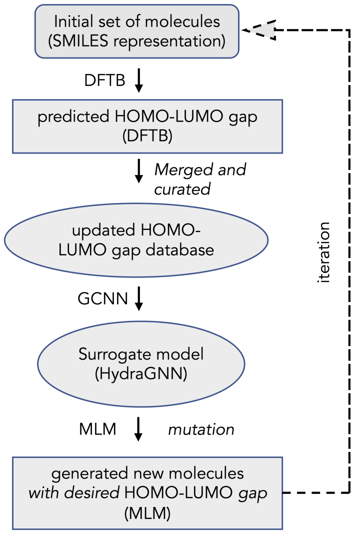
Initial set of molecules (SMILES representation)
DFTB
predicted HOMO-LUMO gap (DFTB)
Merged and curated
updated HOMO-LUMO gap database
GCNN
Surrogate model (HydraGNN)
MLM
mutation
generated new molecules with desired HOMO-LUMO gap (MLM)
iteration

Fig. 1. Computational Workflow for Molecular Design.

DFTB To generate on-the-fly HOMO-LUMO gap predictions, we employ the density-functional tight-binding (DFTB) method. DFTB is a fast and efficient quantum mechanical simulation technique that is implemented e.g. in the DFTB+ program package [23], as well as in other popular quantum chemistry packages. The required runs are performed in two stages. In the first stage we generate the data for predicted HOMO-LUMO gaps from a Simplified Molecular Input Line Entry System (SMILES) string [24] representing the molecular structure. If the structures are in PDB format they can be easily transformed into SMILES representation using the RDKit software [25]. All of the results are merged into a single file and curated for the GCNN surrogate operation.

GCNN surrogate To train a surrogate model for HOMO-LUMO gap prediction, we use the multi-headed HydraGNN package developed earlier by some of our team [18, 19]. This surrogate, allowing multi-headed output, is ideally suited for the simultaneous prediction of multiple important molecular characteristics, such as electronic properties and synthesizability scoring. Training of HydraGNN can be performed on multiple GPU nodes. The trained surrogate model is then used to generate hydra score for subsequent molecule generation operation. Details on the HydraGNN surrogate are given below.

Molecule generation using MLM In our last step we use the masked language model (MLM) to generate novel molecules based on the surrogate model. The MLM was trained following previously reported work [21] on the Enamine *REAL* database [26]. These new molecules are then used to validate the surrogate-predicted HOMO-LUMO gaps using the DFTB method.

3 The DFTB Method

The density-functional tight-binding (DFTB) method [27, 28, 29] is an approximation to traditional density functional theory (DFT) [30], roughly 2-3 orders of magnitude faster yet providing detailed electronic structure by solving Kohn-Sham equations using a parameterized Hamiltonian. DFTB methods can be employed in simulations of processes that involve chemical reactions, electron excitation, electron transfer, and mass and ion transport for systems containing several tens of thousands of atoms. Linear scaling algorithms exist [31, 32] and have been developed and applied to systems as large as 100 million atoms [33].

One of the key features of DFTB is the use of a two-center approximation [27], which requires the pairwise generation of Hamiltonian element parameters and repulsive potentials. The Foulkes-Haydock approach to the expansion of electron density in a Taylor series around a reference density [28] gives rise to a hierarchical family of DFTB flavors, starting with the simplest version DFTB1 [27] which is accurate to a first-order expansion, to the most involved DFTB3 flavor which contains the full third-order terms [29]. All DFTB flavors can be cast into a spin-dependent formalism using on-site spin coupling terms [34]. Moreover, a long-range corrected version of the second-order DFTB2 flavor has been developed [35] in order to overcome the infamous self-interaction error inherent to conventional DFT and DFTB methods. In addition, a variety of ad-hoc charge-charge interaction damping and dispersion interactions have been introduced that can be added to account for potential deficiencies in any of the DFTB flavors [36]. However, the performance of any of these resultant DFTB flavors are strongly dependent on the respective optimized electronic Hamiltonian parameters and repulsive potentials for the chemical element combination in question which need to be optimized for any of the DFTB flavors individually.

High-quality DFTB parameters allow a near-1:1 reproduction of molecular orbital (MO) electronic energy levels for molecules, as well as valence and conduction band structures for bulk solid materials relative to DFT [37]. Repulsive potentials can be optimized such that DFTB calculations are able to reproduce DFT geometrical parameters [29] as well as vibrational [38] and optical [39] spectra. In terms of computational accuracy and efficiency, DFTB is settled in between traditional DFT methods and classical force field approaches, although higher accuracy than DFT can be achieved at times when empirical data is employed in the parameterization, for instance for the prediction of band gaps [40].

In our automated workflow, the DFTB calculations are run as follows. A SMILES string corresponding to an arbitrary molecule is internally converted into a set of Cartesian coordinates that are handed to the DFTB+ code as input, along with the selected level of theory. A python script is started to perform the DFTB calculation via the "atomic simulation environment" (ASE) [41], which allows for software-agnostic electronic structure calculations. The DFTB calculation itself in our case is a geometry optimization with user-controllable convergence criteria. The geometry optimization is necessary since the HOMO-LUMO gap is sensitive to the selected geometry. Ideally, the global minimum energy structure should be identified since it is the molecular conformation likely adopted by the molecule in question. At the end of the calculation, the value of the HOMO-LUMO gap is collected from the output corresponding to the converged molecular structure. There are two shortfalls that may occur in the DFTB calculation step: i) failure to achieve self-consistent-charge convergence of the atomic charges in DFTB density calculations, and ii) failure to achieve a converged minimum energy optimized structure. Our algorithm automatically detects those cases and removes them from the database of molecules.

In the presented proof-of-principles applications, we selected the DFTB3 method [29] with the associated, so-called 3ob parameters [42] for the calculation of MO energies and molecular geometries.

4 Surrogate Models: Graph Convolutional Neural Networks

GCNNs are a class of deep learning models that can operate directly on graphs. Graphs, $G = (V, \mathcal{E})$, are data structures that consist of a finite set of nodes V and edges $\mathcal{E}$ connecting the nodes with their neighbour nodes. For example, an edge, $e = (u, v) \in \mathcal{E}$, connects node $u \in V$ with its neighbour $v \in V$.

Representing molecules in the form of graph is natural. The atoms can be viewed as nodes and chemical bonds as edges of the graph as in Fig. 2. Each node in graph is represented by a nodal feature vector such as atomic features in molecules and potentially label properties in the node-level tasks. For edges, in addition to representing the connectivity of nodes in graph, they can also have edge features such as the chemical bound types. Each graph can have global graph-level properties such as the HOMO-LUMO gap for a specific molecular graph.

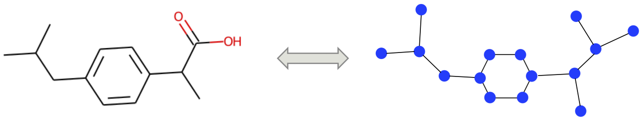

Fig. 2. Graph representation of a molecule.

GCNNs employ a message-passing framework to represent the interactions of nodes, the centerpiece of which is graph convolutional (GC) layers. In each GC layer, messaging passing is performed sequentially with three steps, i.e., message preparation, aggregation, and nodal state update. In message preparation, each node wraps a message, e.g., a vector, based on their current state with a message function and then sends out the message through edges to their neighbours. In message aggregation, the nodes collect the messages received from their neighbours and aggregate the messages with an aggregation function. The aggregated message is then used to update the node together with its current state via an update function. After the message passing in one GC layer, all nodal states are updated with information gathered from their immediate neighbour nodes. With multiple GC layers stacked together, messages are passed to nodes that are further away. In molecules, a single GC layer can be used to approximate pair-wise atomic interactions, while the many-body effects are implicitly represented by stacking many GC layers together. A variety of GCNNs have been developed to better represent the atomic systems, such as crystal GCNN (CGCNN) [43] for crystalline materials, MEGNet for both molecules and crystals [44], as well as ALIGNN [45] that adds the bond angles in the model. More details about GCNNs can be found in [18].

In this work, the GCNN model implemented in HydraGNN [18][19] is used as a surrogate model for HOMO-LUMO gap property. The GC layer used is principal neighborhood aggregation (PNA) [46]. HydraGNN is an open source software built on `PyTorch` [47, 48] and `PyTorch Geometric` [49, 50] library. It is capable of multitask prediction of hybrid node-level and graph-level properties in large batch of graphs with variable number of nodes. HydraGNN utilizes ADIOS [51], a data framework suitable for high performance computing, for data loading in order to handle large graph data set. It has been tested on multiple systems including ORNL's Summit and NERSC's Perlmutter with data sets larger than 100 GB.

Model architecture and training on Summit The GCNN model used in the work consists of three parts, i.e., PNA convolutional layers, batch normalization layers, and fully connected layers. The model starts with a stack of six PNA layers of size 55, each of which is followed by a batch normalization layer, then goes through a global mean pooling layer, and ends with three fully connected layers with 100, 50, 25 neurons, respectively. ReLU activation functions are used.

The DFTB data set consists of 95k molecules with HOMO-LUMO gap generated using the method in Section 3. Ninety percent of the data set is used for model training and the other 10% is split equally for model validation and testing. The AdamW optimizer is used with the learning rate of 10^{-3} and the batch size of 64. The model is trained for 200 epochs and the final test MAE for HOMO-LUMO gap is 0.12 eV.

5 Molecule Generation: Masked Language Model

Research in natural language processing (NLP) has provided strategies to leverage large amounts of unlabeled data to train generalizable language models for text generation and prediction [52]. Masked language models are typically developed using two distinct stages, known as pre-training and fine-tuning. Pre-training is completely unsupervised (i.e., doesn't require any manual labeling) and, therefore, can be performed on very large data sets. In fine-tuning, the pre-trained model is further trained for a specific task (e.g., document classification) using a labeled data set that is typically much smaller. Using this two-stage approach, language models have achieved state-of-the-art results for a range of NLP tasks for multiple application areas [52, 53].

Pre-training language models on text sequences can be accomplished using a combination of tokenization and mask prediction. In tokenization, commonly occurring sequences are used to generate a vocabulary for the model [54, 55]. This vocabulary is used to map text to a sequence of integer token IDs which are used as input to the model. For mask prediction, tokens are randomly masked and the model is trained to reproduce the original sequence based on context. Therefore, for a given masked token, the model predicts a probability that each token in the vocabulary will occur at that location.

Advances in language models can be directly applied to molecular structures by using the Simplified Molecular Input Line Entry System (SMILES) text representation [24]. Using a SMILES string, atoms and bonds for a given molecule are converted to a sequence of characters. For example, benzene is given by c1ccccc1, where c represents an individual aromatic carbon atom and 1 represents the start and end of a ring. Similar to traditional text applications, tokenization can be used to split up a given molecule into commonly occurring subsequences [54, 55]. Mask prediction during pre-training then proceeds with the model learning to predict chemical structure based on context.

In our previous work [21, 56], we proposed a strategy to use pre-trained models to generate new molecular structures. Similar to pre-training, a given molecule, represented as a SMILES string, is tokenized and randomly masked. The model predictions are then sampled to generate a set of mutations to the original molecule as shown in Figure 5. Therefore, in combination with the scoring provided by the surrogate model, the language model can be used to iteratively generate new molecules to search chemical space for a given optimization task. Most notably, the MLM can generate molecules that are much larger in the number of atoms relative to the original set of molecules. This is particularly important when considering the rapidly increasing chemical space with molecular size.

Pre-training on Summit Following our previous work [21], we leveraged the Enamine *REAL* database [26] as a starting point for language model training. We then augmented the data set using a previously trained language model [21] to include approximately $3.6 \cdot 10^{10}$ molecules. A WordPiece tokenizer [54, 55] was then trained using the full data set. As described in [21], we used data parallelism with DeepSpeed's fused LAMB optimizer for mask prediction training on $3 \cdot 10^{9}$ molecules on the Summit supercomputer at the Oak Ridge Leadership Computing Facility using 1000 nodes (6 Nvidia 16 GB V100 GPUs per node). The data set was evenly partitioned with $5 \cdot 10^{5}$ molecules for each GPU. Pre-training was performed for 7 epochs using a batch size of 80 molecules with 3 gradient accumulation steps per GPU. The model is available for download at https://huggingface.co/mossaic-candle/adaptive-lm-molecules and can be used with the Hugging Face transformers library [57].

6 Application: Minimizing the HOMO-LUMO gap

In this Section we describe our successful implementation of the end-to-end workflow for accelerating the molecular design process by coupling the approximate quantum chemical methods (i.e. DFTB), surrogate GCNN model, and the generative MLM used as mutation operator for the molecule generation, for minimizing the HOMO-LUMO gap as the target property.

For future reference, we label the 95k molecular compounds that were contained in QM9 and successfully processed by the DFTB calculations to generate their optimized molecular geometries and associated HOMO-LUMO gaps as the "original data set". As described in Section 4, this data set was split into training and test data sets.

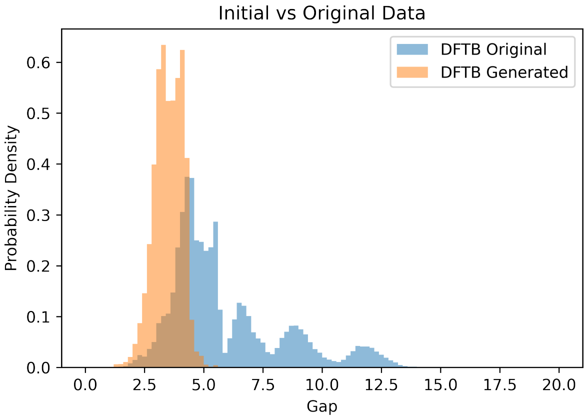

Fig. 3. Comparison of the DFTB-predicted HOMO-LUMO gap is shown for molecules contained in the original data set and for newly generated molecules. The population of molecules clearly shifted towards the target low values of the molecular HOMO-LUMO gap.

Figure 3 shows the comparison of the HOMO-LUMO gap as predicted by DFTB for the molecules contained in the original data set (blue) as well as for the newly generated ones (orange). At first glance it becomes apparent that the generated molecules possess much lower HOMO-LUMO gap values, following the user-defined constraint. It further appears that, while the molecules in the original data set show a multi-modal distribution of their HOMO-LUMO gaps corresponding to the different molecular classes (aliphatic molecules > olefinic molecules > conjugated molecules > molecules with double bonds and strained rings), the new predicted molecules have a more concentrated distribution centered near the 3–3.5 eV (see Figs. 4 and 5).

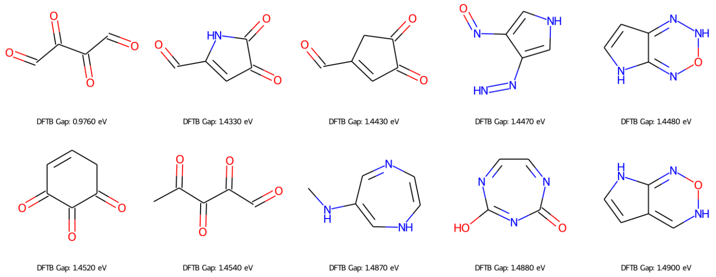

Fig. 4. Selected molecules with HOMO-LUMO gaps < 1.7 eV that are already contained in the original QM9 data set. None of these molecules contains more than 9 "heavy" (C,N,O) chemical elements.

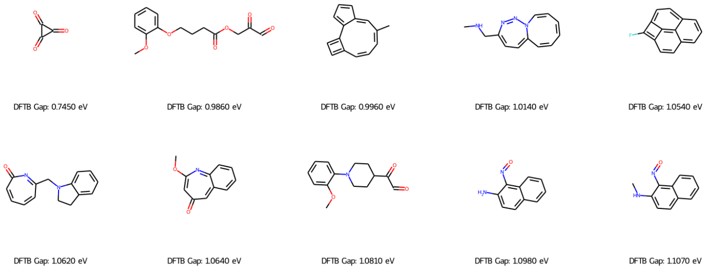

Fig. 5. Selected molecules with HOMO-LUMO gaps < 1.7 eV contained in the generated data set. Many of the generated molecules with low HOMO-LUMO gaps contain more than 9 atoms.

The latter figure likewise indicates that most of the newly generated molecules contain more than 9 atoms, and that such structures were not at all included in the original training set of the GCNN surrogate. We do note that the generated molecules show a tendency towards including small, strained rings with double bonds included, such as three- and four-membered rings that are often fused to larger rings. Such molecules derive their small HOMO-LUMO gaps from their ring strains which typically pushes the HOMO levels up and the LUMO levels down, reducing the HOMO-LUMO gap [39], but at the same time increasing their chemical reactivity and the difficulty of synthesis.

Synthesizability is a quantity that can be calculated following the method as mentioned by Etrl et al. [58]. In this empirical technique the synthetic accessibility is estimated by coupling molecular complexity and molecular fragment effects as analyzed by processing a large collection of chemical structures that have already been synthesized before. This technique therefore contains at some level historical data on the synthesizability of a large collection of molecules. The method is then tested and validated by comparing with 'ease of synthesis' ranks. These values were gathered by domain expert chemists. The agreement between score and the ranks are generally very good. This technique also provides a good way to estimate the synthesizability for large molecules that have never been synthesized before.

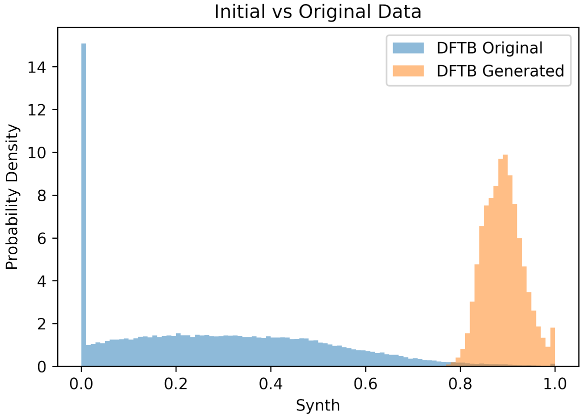

Fig. 6. Comparison of the synthesizability score for molecules contained in the original data set and for newly generated molecules. As the HOMO-LUMO gap decreases, the synthesizability score increases. For discussion, see text.

Fig. 6 indicates that, in contrast to the molecules in the original data set, the newly generated ones have much higher synthesizability scores, inline with the above-mentioned fact that they feature strained molecular structures and conjugated bonds with heteroatoms (such as carbonyl groups) that can be seen as easily decomposable in potentially violent explosions such as the first molecule shown in Fig. 5. This indicates the necessity that a molecular design scheme focusing on a single target property such as low HOMO-LUMO gap may not be very successful in delivering real-world solutions, and that a multi-objective search needs to be performed instead that will result in a Pareto-optimal set of molecules, combining a number of design targets such as low HOMO-LUMO gaps paired with low values of synthesizability.

Fig. 7 gives an impression of the Mean Absolute Error (MAE) of the GCNN surrogate predictions against DFTB-calculated HOMO-LUMO gaps for the molecules

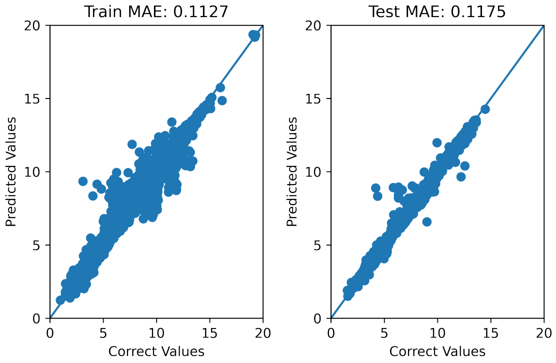

Fig. 7. Comparison of GCNN-predicted HOMO-LUMO gaps vs DFTB-computed ones ("Correct Values") for the training (left) and the test (right) data set of molecules contained in the QM9 database.

contained in the original data set, separately for molecules of the training set and the test set. The MAE is nearly identical in both close to 0.11 eV, indicating an excellent performance of the HydraGNN surrogate for this task, and the homogeneity of the data set.

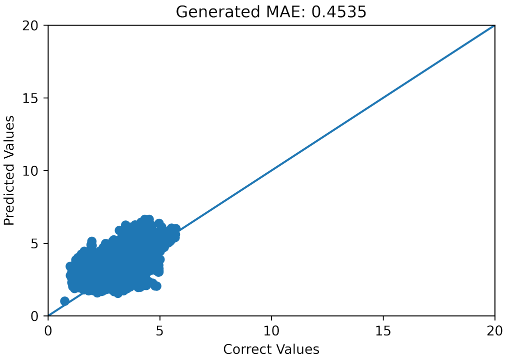

Fig. 8. Comparison of GCNN-predicted HOMO-LUMO gaps vs DFTB-computed ones ("Correct Values") for the MLM-generated molecules with smaller HOMO-LUMO gaps.

Such good agreement between surrogate- and DFTB-predicted HOMO-LUMO gaps is lost when considering the situation of the newly generated molecules, shown in Fig. 8. The MAE is increased to 0.45 eV as the now larger molecules, not contained in the original data set, lead to poorer performance of the HydraGNN surrogate which was only trained on the QM9 database containing molecules with up to 9 heavy chemical elements. This fact indicates the usefulness to perform adaptive surrogate training in iterative molecular design approaches, as indicated in Fig. 1 by the dashed arrow labeled “iteration”. Nevertheless, it should be noted that the GCNN surrogate model, even without iterative improvements, was able to provide sufficiently reliable guidance to the MLM for generating new molecular structures with reduced HOMO-LUMO gap.

For completeness, and to give an impression of the chemical variability contained in the designed novel organic molecules with low HOMO-LUMO gap, we show their molecular structures in the associated supplementary information (SI). Fig. S1 shows all molecules with HOMO-LUMO gaps < 1.7 eV contained in the original data set (43) and Fig. S2 shows those in the data set of generated, novel molecules with the same HOMO-LUMO gap threshold (384).

7 Conclusions and Future Work

The current proof-of-principle shows that the combination of semiempirical quantum chemical electronic structure theory, GCNN surrogate, and MLM generative model succeeds to predict a plethora of novel molecular compounds with desirable optical properties, in this case HOMO-LUMO gaps that are as small as possible. First we wish to mention how well our multi-headed GCNN surrogate reproduced HOMO-LUMO gaps, the MAE was 0.11 eV for molecules both in the training as well as test set (both part of the QM9 database). This accuracy is similar to the one recently reported by Lilienfeld et al. [13] and is comparable, or even exceeds, the error bars that can be expected from traditional DFT methods in their prediction of excitation energies, and thus lays the foundation for an inverse design workflow focusing on HOMO-LUMO gaps. The surrogate lost accuracy when trying to predict HOMO-LUMO gaps of larger molecules that were not part of the training set, which indicates the necessity of adaptive surrogate training schemes in iterative molecular design. Our computational workflow for accelerated molecular design using quantum chemical simulations and deep learning models already possesses these capabilities, and we will exploit them in future applications.

Nevertheless, even in a single iteration, our workflow succeeded in predicting a significantly large number of molecules with very small HOMO-LUMO gaps < 1.7 eV (384) which is much larger than the fraction of such molecules contained in the entire QM9-based original data set containing 95k molecules (43). This constitutes proof-of-principle of our combined surrogate/generative model approach, even factoring in that the MLM was not even trained on molecules with particular optical properties.

The newly generated molecules have the caveat that their synthesizability scores are high, meaning that they are not easily to synthesize or are unstable for other reasons, which underlines the necessity of multi-objective optimizations. In future works, we will test the capability to perform adaptive training built into our workflow, and incorporate advanced property predictions such as the prediction of electronic excitation energies and oscillator strengths for real-world design applications targeting photoactive molecules and molecular dyes. Our workflow will be generally applicable to any type of molecular properties that can be predicted by quantum chemical electronic structure programs, such as molecular energies, electronic and magnetic properties, vibrational properties, and optical properties in general.

Acknowledgements

This work was supported in part by the Office of Science of the Department of Energy and by the Laboratory Directed Research and Development (LDRD) Program of Oak Ridge National Laboratory. This research is sponsored by the Artificial Intelligence Initiative as part of the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US Department of Energy under contract DE-AC05-00OR22725. The research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. This work used resources of the Oak Ridge Leadership Computing Facility, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

References

[1] Basic Energy Sciences Advisory Committee et al. “Directing Matter and Energy: Five Challenges for Science and the Imagination”. *US Department of Energy: Washington, DC* (2007).

[2] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. “Inverse molecular design using machine learning: Generative models for matter engineering”. *Science* 361.6400 (2018), pp. 360–365.

[3] Andrew E. Blanchard, Christopher Stanley, and Debsindhu Bhowmik. “Using GANs with adaptive training data to search for new molecules”. *J Cheminform* 13.14 (2021), pp. 1–8. DOI: 10.1186/s13321-021-00494-3.

[4] Wenbo Sun et al. “Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials”. *Sci. Adv.* 5.11 (2019), eaay4275.

[5] Pavlo O Dral and Mario Barbatti. “Molecular excited states through a machine learning lens”. *Nat. Rev. Chem.* 5.6 (2021), pp. 388–405.

[6] Alex Zhavoronkov. “Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry”. *Mol. Pharm.* 15.10 (2018), pp. 4311–4313.

[7] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. “Drug discovery with explainable artificial intelligence”. *Nat. Mach. Intell* 2.10 (2020), pp. 573–584.

[8] Debsindhu Bhowmik et al. "Deep clustering of protein folding simulations". *JBMC Bioinformatics* 19.484 (2018), pp. 47–58. DOI: h10.1186/s12859-018-2507-5.

[9] Ya Zhuo and Jakoah Brgoch. "Opportunities for next-generation luminescent materials through artificial intelligence". *J. Phys. Chem. Lett.* 12.2 (2021), pp. 764–772.

[10] Cheng-Wei Ju et al. "Machine learning enables highly accurate predictions of photophysical properties of organic fluorescent materials: Emission wavelengths and quantum yields". *J. Chem. Inf. Model* 61.3 (2021), pp. 1053–1065.

[11] Atanu Acharya et al. "Supercomputer-based ensemble docking drug discovery pipeline with application to COVID-19". *J. Chem. Inf. Model* 60.12 (2020), pp. 5832–5852.

[12] Nastaran Meftahi et al. "Machine learning property prediction for organic photovoltaic devices". *Npj Comput. Mater* 6.1 (2020), pp. 1–8.

[13] Bernard Mazouin, Alexandre Alain Schöpfer, and O Anatole von Lilienfeld. "Selected Machine Learning of HOMO-LUMO gaps with Improved Data-Efficiency". *arXiv preprint arXiv:2110.02596* (2021).

[14] Kerstin Andersson, Per-Åke Malmqvist, and Björn O Roos. "Second-order perturbation theory with a complete active space self-consistent field reference function". *J. Chem. Phys* 96.2 (1992), pp. 1218–1226.

[15] Celestino Angeli et al. "Introduction of n-electron valence states for multireference perturbation theory". *J. Chem. Phys* 114.23 (2001), pp. 10252–10264.

[16] Silvana Botti et al. "Time-dependent density-functional theory for extended systems". *Rep. Prog. Phys.* 70.3 (2007), p. 357.

[17] Monja Sokolov et al. "Analytical Time-Dependent Long-Range Corrected Density Functional Tight Binding (TD-LC-DFTB) Gradients in DFTB+: Implementation and Benchmark for Excited-State Geometries and Transition Energies". *J. Chem. Theory Comput.* 17.4 (2021), pp. 2266–2282.

[18] M. Lupo Pasini et al. "Multi-task graph neural networks for simultaneous prediction of global and atomic properties in ferromagnetic systems". *Mach. learn.: sci. technol.* 3.2 (2022), p. 025007. DOI: 10.1088/2632-2153/ac6a51. URL: https://doi.org/10.1088/2632-2153/ac6a51.

[19] Massimiliano Lupo Pasini et al. *HydraGNN*. [Computer Software] https://doi.org/10.11578/dc.20211019.2. 2021. DOI: 10.11578/dc.20211019.2. URL: https://github.com/ORNL/HydraGNN.

[20] Benhao Li, Mengyao Zhao, and Fan Zhang. "Rational design of near-infrared-II organic molecular dyes for bioimaging and biosensing". *ACS Mater. Lett.* 2.8 (2020), pp. 905–917.

[21] Andrew E. Blanchard et al. "Language Models for the Prediction of SARS-CoV-2 Inhibitors". *bioRxiv* (2021). DOI: 10.1101/2021.12.10.471928. URL: https://www.biorxiv.org/content/10.1101/2021.12.10.471928v1.

[22] Raghunathan Ramakrishnan et al. "Quantum chemistry structures and properties of 134 kilo molecules". *Sci. Data* 1.1 (2014), pp. 1–7.

[23] Ben Hourahine et al. "DFTB+, a software package for efficient approximate density functional theory based atomistic simulations". *J. Chem. Phys.* 152.12 (2020), p. 124101.

[24] David Weininger. "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". *J. Chem. Inf. Comput. Sci.* 28 (1998), pp. 31–36. DOI: https://doi.org/10.1021/ci00057a005.

[25] *RDKit: Open-source cheminformatics.* 2022. URL: https://www.rdkit.org.

[26] *Enamine* REAL *Database.* https://enamine.net/compound-collections/real-compounds/real-database. Accessed: 2020-04-01, through https://virtual-flow.org/.

[27] D. Porezag et al. "Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon". *Phys. Rev. B* 51.19 (May 1995), pp. 12947–12957. DOI: 10.1103/PhysRevB.51.12947. URL: https://link.aps.org/doi/10.1103/PhysRevB.51.12947.

[28] M. Elstner et al. "Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties". *Phys. Rev. B* 58.11 (Sept. 1998), pp. 7260–7268. DOI: 10.1103/PhysRevB.58.7260. URL: https://link.aps.org/doi/10.1103/PhysRevB.58.7260.

[29] Michael Gaus, Qiang Cui, and Marcus Elstner. "DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)". *J. Chem. Theory Comput.* 7.4 (2011), pp. 931–948. ISSN: 1549-9618, 1549-9626. DOI: 10.1021/ct100684s. URL: https://pubs.acs.org/doi/10.1021/ct100684s.

[30] Robert O Jones. "Density functional theory: Its origins, rise to prominence, and future". *Rev. Mod. Phys.* 87.3 (2015), p. 897.

[31] Yoshio Nishimoto, Dmitri G. Fedorov, and Stephan Irle. "Density-Functional Tight-Binding Combined with the Fragment Molecular Orbital Method". *J. Chem. Theory Comput.* 10.11 (2014), pp. 4801–4812. ISSN: 1549-9618. DOI: 10.1021/ct500489d. URL: https://pubs.acs.org/doi/10.1021/ct500489d.

[32] Yoshifumi Nishimura and Hiromi Nakai. "Dcdftbmd: Divide-and-Conquer Density Functional Tight-Binding Program for Huge-System Quantum Mechanical Molecular Dynamics Simulations". *J. Comput. Chem.* 40.15 (2019), pp. 1538–1549. ISSN: 1096-987X. DOI: 10.1002/jcc.25804. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.25804.

[33] Yoshifumi Nishimura and Hiromi Nakai. "Quantum Chemical Calculations for up to One Hundred Million Atoms Using DCDFTBMD Code on Supercomputer Fugaku". *Chem. Lett.* 50.8 (2021), pp. 1546–1550.

[34] Thomas Frauenheim et al. "Atomistic simulations of complex materials: ground-state and excited-state properties". *J. Phys. Condens. Matter* 14.11 (2002), p. 3015.

[35] Vitalij Lutsker, Balint Aradi, and Thomas A Niehaus. "Implementation and benchmark of a long-range corrected functional in the density functional based tight-binding method". *J. Chem. Phys.* 143.18 (2015), p. 184107.

[36] Jan Rezac. “Empirical self-consistent correction for the description of hydrogen bonds in DFTB3”. *J. Chem. Theory Comput.* 13.10 (2017), pp. 4804–4817.

[37] Qiang Cui and Marcus Elstner. “Density functional tight binding: values of semi-empirical methods in an ab initio era”. *Phys. Chem. Chem. Phys.* 16.28 (2014), pp. 14368–14377.

[38] Yoshio Nishimoto and Stephan Irle. “Quantum chemical prediction of vibrational spectra of large molecular systems with radical or metallic electronic structure”. *Chem. Phys. Lett.* 667 (2017), pp. 317–321.

[39] Cristopher Camacho et al. “Origin of the size-dependent fluorescence blueshift in [n] cycloparaphenylenes”. *Chem. Sci.* 4.1 (2013), pp. 187–195.

[40] Chien-Pin Chou et al. “Automatized parameterization of DFTB using particle swarm optimization”. *J. Chem. Theory Comput.* 12.1 (2016), pp. 53–64.

[41] Ask Hjorth Larsen et al. “The atomic simulation environmenta Python library for working with atoms”. *J. Phys. Condens. Matter* 29.27 (2017), p. 273002.

[42] Maximilian Kubillus et al. “Parameterization of the DFTB3 Method for Br, Ca, Cl, F, I, K, and Na in Organic and Biological Systems”. *J. Chem. Theory Comput.* 11.1 (2015), pp. 332–342. ISSN: 1549-9618. DOI: 10.1021/ct5009137. URL: https://doi.org/10.1021/ct5009137 (visited on 06/03/2021).

[43] T. Xie and J. C. Grossman. “Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties”. *Phys. Rev. Lett.* 120.14 (Apr. 2018), p. 145301. DOI: 10.1103/PhysRevLett.120.145301. URL: https://link.aps.org/doi/10.1103/PhysRevLett.120.145301.

[44] Chi Chen et al. “Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals”. *Chem. Mater.* 31.9 (2019), pp. 3564–3572. DOI: 10.1021/acs.chemmater.9b01294. eprint: https://doi.org/10.1021/acs.chemmater.9b01294. URL: https://doi.org/10.1021/acs.chemmater.9b01294.

[45] Kamal Choudhary and Brian DeCost. “Atomistic Line Graph Neural Network for improved materials property predictions”. *Npj Comput. Mater.* 7.1 (2021), pp. 1–8.

[46] G. Corso et al. “Principal Neighbourhood Aggregation for Graph Nets”. en. *arXiv:2004.05718 [cs, stat]* (Dec. 2020). arXiv: 2004.05718. URL: http://arxiv.org/abs/2004.05718 (visited on 02/21/2021).

[47] A. Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. *Adv Neural Inf Process Syst. 32.* Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[48] *PyTorch.* https://pytorch.org/docs/stable/index.html.

[49] M. Fey and J. E. Lenssen. “Fast Graph Representation Learning with PyTorch Geometric”. *ICLR Workshop on Representation Learning on Graphs and Manifolds*. 2019.

[50] *PyTorch Geometric*. https://pytorch-geometric.readthedocs.io/en/latest/.

[51] William F. Godoy et al. “ADIOS 2: The Adaptable Input Output System. A framework for high-performance data management”. *SoftwareX* 12 (2020), p. 100561. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2020.100561. URL: https://www.sciencedirect.com/science/article/pii/S2352711019302560.

[52] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* 1.Mlm (2019), pp. 4171–4186. arXiv: 1810.04805.

[53] Yu Gu et al. “Domain-specific language model pretraining for biomedical natural language processing”. *arXiv* (2020). ISSN: 23318422. URL: https://arxiv.org/abs/2007.15779.

[54] Mike Schuster and Kaisuke Nakajima. “Japanese and Korean voice search”. *2012 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*. 2012, pp. 5149–5152. DOI: 10.1109/ICASSP.2012.6289079.

[55] Yonghui Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation” (2016). arXiv: 1609.08144. URL: http://arxiv.org/abs/1609.08144.

[56] Andrew E. Blanchard et al. “Automating Genetic Algorithm Mutations for Molecules Using a Masked Language Model”. *IEEE Trans. Evol. Comput.* (2022). DOI: 10.1109/TEVC.2022.3144045.

[57] Thomas Wolf et al. “Transformers: State-of-the-Art Natural Language Processing”. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[58] P. Ertl and Schuffenhauer. “A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions.” *J Cheminform* 1.8 (2009). DOI: https://doi.org/10.1186/1758-2946-1-8.